NMSA407: Linear Regression

Winter Term 2019/2020

General Instructions & Homework Assignment no. 2

(Submission Deadline: Exercise class no. 8)

ℹ General Instructions

- ❑ The homework assignment can be carried out in a group of 1 – 3 students (three students per each group are recommended). The groups are not required to be the same as those in you had for the elaboration of the first homework assignment.

- ❑ Each group is required to submit a well-written `pdf` document created in LaTeX. All content should be nicely formatted in a human-readable form (a format analogous to a bachelor thesis). Computer codes or originally formatted computer outputs should not appear in the document.

- ❑ The document can be submitted either in English or Czech/Slovak (Czech and Slovak are allowed to be used within one document). Do not use English and Czech/Slovak in one document. The language used in the plot labels and figures should correspond with the language used in the main document.

- ❑ The submitted document must contain the names of all members of the group and their exercise group identification in the header on the first page.

- ❑ All statistical tests are performed at 5 % significance level, confidence intervals should be all with the 95% coverage.

- ❑ **DEADLINES and FORM OF DELIVERY:**

Tuesday's session	(*Matúš Maciak*)	Due on 19/11 (13:50)	PRINTED ON PAPER
Tuesday's session	(*Stanislav Nagy*)	Due on 19/11 (14:00)	PRINTED ON PAPER
Thursday's session	(*Matúš Maciak*)	Due on 21/11 (10:30)	PRINTED ON PAPER

- ❑ **REVISIONS:**

  Revised homework solutions have to be accompanied by a letter with a list of all significant changes made to the originally submitted document (e.g. *"A new paragraph with the description of Figure 1 was added on page 8;"*, *"Formatting of the tables throughout the document was improved as suggested;"*, or *"Section 7 of the document was rewritten completely."*). In case when some of the suggested improvements are not carried out in the revision, reasons for not including them should be explained in the letter.

☞ Data Description

A total of 149 urological patients undertook a surgery in order to remove kidney stones. The surgeries took all place at a university hospital in Banská Bystrica in 2014–2016. From the technological point of view, the surgery can be either invasive (holmium based treatment using a flexible YAG Laser) or noninvasive (an ultrasonic laser PEK). Each patient undertook exactly one surgery while the surgery type was decided for each patient at random. The information about the surgery type is part of the dataset. In addition, for each patient there is also some additional patient's specific information recorded in the data (e.g. gender, age, surgery time, surgeon who performed the surgery, etc.).

- ❑ The datafile (an `RData` file) can be downloaded by clicking on the link hw2_2019.RData or by downloading the corresponding file from the central webpage of the NMSA 407 exercise classes.

- ❑ The downloaded data file can be loaded from the working directory (check/set your working directory using the commands `getwd()` and `setwd()`) into the `R` working environment by using the following command:

  ```
  > load("hw2_2019.RData")
  ```

  The `R` variable with the dataset is called `data`.

- ❑ The dataset contains 149 observations and 8 covariates.

  a) `gender` - patient's gender (`male` or `female`);

  b) `flexPek` - two level factor covariate to distinguish for the noninvasive surgery (`pek`) or an invasive surgery (`flex`);

  c) `surgeon` - four-value covariate to identify the surgeon who performed the surgery;

  d) `size` - numerical covariate which stands for the overall size of the kidney stone(s) given in a diameter [mm];

  e) `SFR` - indicator covariate to express the surgery result: Stone Free Rate (SFR) equals to one if there were no kidney stones remaining and it is equal to zero otherwise;

  e) `time` - the overall time the surgery took place [min];

  f) `intervention` - integer covariate which stands for the number of required interventions during the surgery – if the surgery goes well no interventions are expected;

  g) `age` - patient's age given in years.

The general theme of this homework is the exploration of the dependence of the surgery time (variable `time`) on the size of the kidney stone (variable `size`) when taken the remaining covariates into consideration.

✍ Homework 2 Assignments

Part 1:
Describe the data — present a table of descriptive statistics and suitable figures, and discuss the particularities of the dataset.

Part 2:
Fit a linear model `m1` with the surgery time as a response and other variables as explanatory variables. Do not include interactions yet, but make sure that the model is parametrized so that all regression coefficients (including the intercept) have meaningful interpretations. Create a table which summarizes the important features of model `m1`. In the table, all important results from `summary(m1)` must be given, and explained in detail. Where it makes good practical sense, confidence intervals for regression coefficients, and p-values of tests on regression coefficients should be included and interpreted.

Part 3:
In words interpret each regression coefficient (or a group of coefficients if they all describe a similar quantity). Also a non-statistician should be able to understand the meaning of the model. Discuss whether the model is suitable for predicting the surgery time based on the considered predictors.

Part 4:
Include the basic residual plots for model `m1`. Comment on the validity of the assumptions of a classical normal linear model. Do not perform formal statistical tests.

Part 5:
Provide a table of summary statistics of `time` when differentiated by the `surgeon` covariate (one column per surgeon). We suspect that surgeon no. 3 needs more time to perform a surgery than surgeon no. 4.

1. Provide an estimate (based on `m1`), including the standard error and a $95\,\%$ confidence interval, for the difference between the time needed by these two surgeons. Explain the effect and describe which approach (reference to the lecture notes, etc.) you are using to arrive at the final numbers.
2. Perform a statistical test to decide whether surgeon no. 3 is slower than surgeon no. 4. As always, specify (mathematically) the statistical hypothesis, provide the value of the test statistic, p-value (and how it is computed) and your conclusion expressed in words understandable by a non-statistician. Is it possible that the detected differences can be explained by the fact that surgeon no. 3 had to perform more difficult surgeries (larger kidney stones, more interventions needed, older patients) than surgeon no. 4?
3. Visualize the difference between time needed by surgeons no. 3 and no. 4 using a scatterplot of the surgery time versus the size of the kidney stone based on a subsets of data where you distinguish by different options (symbols, colors, etc.) for the `surgeon` covariate. Add to the plot the fitted regression lines showing the model-based estimated dependence of (mean) time on `size` for considered options of the `surgeon` covariate.

Part 6:
Assume that surgeon no. 2 is going to perform an invasive type of the surgery (flex) on a 45-years-old female patient with the overall size of her kidney stone equal to 22 mm. We are interested in the time needed to complete the surgery. Estimate this time, and provide a $95\,\%$ confidence interval.

Try to find a reasonable solution although only values of some of the variables are specified.

Have a look at the appropriate diagnostic plots and discuss how much trustworthy is the confidence interval that you have just calculated. Do you think that there might be some problem here?

Part 7:
Estimate the expected difference in the time needed to complete the surgery between the two surgery methods pek and flex when both applied by the same surgeon on the same patient. However, the kidney stones sizes are 10 and 20 mm, respectively. Provide a confidence interval for this expected difference.

Part 8:
Modify model m1 by considering either an appropriate transform of the response time, or the logarithmic transformation of size. Compare these alternative models with m1. Which model do you prefer and why? Denote the preferred model by m2.

Part 9:
Extend model m2 so that the variable surgeon possibly modifies the effect of size on time (or on their logarithms). Denote this model by m3 and suppose that it is a useful model.

1. Provide a formal model specification (model formula) for m3 in your report. It is not necessary to include the estimates, or the summary statistics for the regression coefficients.
2. In detail describe the effect of size on time as estimated by model m3.
3. Perform a formal test whether the variable surgeon modifies the effect of size on time, and explain in words its meaning.